Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Genomics 88 (2006) 143–151

GENOMICS

www.elsevier.com/locate/ygeno

An anthropoid-specific segmental duplication on human chromosome 1q22

Vladimir Yu. Kuryshev [a,1], Eugene Vorobyov [b,1], Dorothea Zink [c], Jürgen Schmitz [d], Timofey S. Rozhdestvensky [d], Ewald Münstermann [a], Ute Ernst [a], Ruth Wellenreuther [a], Petra Moosmayer [a], Stephanie Bechtel [a], Ingo Schupp [a], Jürgen Horst [b], Bernhard Korn [c], Annemarie Poustka [a], Stefan Wiemann [a,*]

[a] *Department of Molecular Genome Analysis, DKFZ–German Cancer Research Center, INF 580, D-69120 Heidelberg, Germany*
[b] *Institute of Human Genetics, University of Münster, Vesaliusweg 12-14, D-48149 Münster, Germany*
[c] *Ressourcenzentrum für Genomforschung, RZPD, INF 580, D-69120 Heidelberg, Germany*
[d] *Institute of Experimental Pathology/Molecular Neurobiology (ZMBE), University of Münster, von-Esmarch-Strasse 56, D-48149 Münster, Germany*

Received 9 December 2005; accepted 5 February 2006
Available online 20 March 2006

Abstract

Segmental duplications (SDs) play a key role in genome evolution by providing material for gene diversification and creation of variant or novel functions. They also mediate recombinations, resulting in microdeletions, which have occasionally been associated with human genetic diseases. Here, we present a detailed analysis of a large genomic region (about 240 kb), located on human chromosome 1q22, that contains a tandem SD, SD1q22. This duplication occurred about 37 million years ago in a lineage leading to anthropoid primates, after their separation from prosimians but before the Old and New World monkey split. We reconstructed the hypothetical unduplicated ancestral locus and compared it with the extant SD1q22 region. Our data demonstrate that, as a consequence of the duplication, new anthropoid-specific genetic material has evolved in the resulting paralogous segments. We describe the emergence of two new genes, whose new functions could contribute to the speciation of anthropoid primates. Moreover, we provide detailed information regarding structure and evolution of the SD1q22 region that is a prerequisite for future studies of its anthropoid-specific functions and possible linkage to human genetic disorders.

Keywords: *Alu* repeats; Chimeric gene; Primate evolution; Retroposition

Gene duplication is known to be a major mechanism to create new genetic material with novel functions during evolution [1,2]. Initial analysis of the human genome revealed large segments of nearly identical sequences in particular chromosomal regions. The recent origin of these segments and their abundance (approximately 5% of the total genome) have challenged investigators to elucidate the mechanisms underlying their emergence and their role in primate genome evolution. Some of these duplicated segments have recently been shown to be associated with rapid gene innovation and chromosomal rearrangement in man and the great apes [3,4]. There is evidence for a highly nonrandom chromosomal and genic distribution of recent segmental duplications (SDs), with a likely role in expanding protein diversity [5]. The number of SDs is known to have increased dramatically during the last 35 million years of primate evolution [3]. Many such duplication events coincided with amplification of the *Alu* repeat family. It has, therefore, been suggested that the two phenomena might be causally related [6]. The unexpected finding of such large-scale genomic rearrangements late in primate evolution also raised the possibility that new genes may have been created on a scale that was not previously expected. Moreover, SDs, due to the high sequence similarity of their duplicated segments, have the

Abbreviations: FLAM-C, free left *Alu* monomer of type C; MER, medium reiteration frequency repeat; Mya, million years ago; NWM, New World monkey; OWM, Old World monkey; RT-PCR, reverse transcription-polymerase chain reaction; SD, segmental duplication; SD1q22, segmental duplication on human chromosome 1q22; UTR, untranslated region.

* Corresponding author. Fax: +49 6221 423454.
E-mail address: s.wiemann@dkfz.de (S. Wiemann).

[1] These authors contributed equally to the work.

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.02.002

potential to mediate high-frequency homologous recombination events, many of which have been implicated in human genetic disease [7–10].

During a large-scale annotation of cDNA sequences [11,12] we found a large (about 240 kb) tandem segmental duplication located on human chromosome 1q22 (SD1q22). This SD is not yet fully characterized in the most representative SD-related databases (Segmental Duplication Database, http://humanparalogy.gs.washington.edu, and Human Genome Segmental Duplication Database, http://projects.tcag.ca/humandup). Our preliminary analysis of this SD revealed that both segments exhibit a high level of sequence identity, suggesting its relatively recent origin. As such, SD1q22 may contain new genetic information associated with recent evolutionary events and could be responsible for the speciation of the lineage leading to primates or their suborders or families. In addition, this genomic locus was found to be amplified in cells of several cancer types: ovarian and breast tumors [13], sarcomas [14,15], and hepatocellular carcinomas [16]. This suggests that dosage imbalances of one or more genes in this locus may contribute to the cancer development. Considering the potential importance of SD1q22, we have performed a detailed analysis of the gene content and evolution of this genomic region. We show that the duplication event resulting in formation of SD1q22 occurred about 37 million years ago (Mya) in a lineage leading to anthropoid primates, after their separation from the prosimians but before the Old and New World monkey split. Our data demonstrate that, as a consequence of the duplication, new anthropoid-specific genetic material has evolved in the resulting paralogous segments. We describe the emergence of two new genes, which may have acquired new functions and, thus, contributed to the speciation of the anthropoid primates.

Results and discussion

Tandem segmental duplication on human chromosome 1q22

SD1q22 (chromosome 1: 152330107–152569771) spans about 240 kb and consists of two segments; left segment (LS) and right segment (RS) (proximal and distal to the centromere, respectively) are attached to each other in a “head-to-tail” manner. The LS is almost 135 kb in length (chromosome 1: 152330107–152464858) and the RS spans about 105 kb (chromosome 1: 152464859–152569771). The difference in segment lengths results from postduplication deletions and insertions. No sequences similar to SD1q22 (excluding repeats) could be found elsewhere in the genome, suggesting that this SD is unique.

Sequence analysis of the SD1q22 junction regions revealed the presence of two paralogous FLAM-C type *Alu* repeats surrounding RS (Fig. S1). The left border of LS does not contain a FLAM-C repeat, although it is rich in other *Alu* repeats. This suggests a possible mechanism for the duplication via *Alu*–*Alu*-mediated recombination, which is described as the predominant mechanism of intrachromosomal tandem segmental duplications [17,18]. The absence of the third FLAM-C repeat in the left border of LS could be explained, in turn, by its subsequent deletion, mediated by neighboring *Alu* sequences, or by a gene conversion event, in which one type of *Alu* might have been substituted by a similar repeat of another type [19].

SD1q22 is an anthropoid-specific duplication that occurred about 37 Mya

The high degree of sequence identity observed between LS and RS (95.7 and 93.4% with and without exons, respectively, Table S1) suggests a relatively recent origin of this SD. We found that SD1q22 orthologous loci in the chimpanzee and macaque genomes contain a similar duplication (data not shown). By contrast, the orthologous region is not duplicated in rodents and birds (data not shown). To date the time of duplication more precisely we used the distribution of duplicated and nonduplicated *Alu* elements as a molecular clock. Different types of *Alu* sequences have expanded by successive waves of active expression and retroposition that were limited in time during the evolution of primates [20]. Thus, particular types of *Alu* repeats, which are shared and not shared by the LS and RS, could indicate a certain time period associated with their expansion. We identified 27 transposed *Alu* elements along with their flanking direct repeats (Fig. S2). Six of these *Alu* elements are located in identical positions within both LS and RS. All the shared *Alu* sequences clearly belong to the same *Alu* subtypes. Eleven *Alu* sequences are specific to the RS and 10 are found in LS only, all contain recognizable empty target sites at the corresponding paralogous segment.

The youngest of the shared *Alu* elements belongs to the *AluSp* subfamily and is flanked by 14 bp-long almost identical direct repeats. As shown in Fig. 1, the duplication took place after the *AluSp* element integrated into the original unduplicated chromosomal segment. The common origin of the shared *AluSp* sequences is further supported by five autapomorphic changes, excluding hypermutable CpG sites that are absent in the canonical consensus sequence of *AluSp*. Because the *AluSp* elements were active about 37 Mya [20], the duplication event can also be placed around this time. After the duplication, *Alu* repeats continued to accumulate individually in the duplicons. We detected *AluS* and younger, *AluY,* elements specific for either of the segments but no older *Alu* subfamily members (e.g., from the *AluJ* subfamily or *Alu* monomers). Therefore, both shared and unique *Alu* insertions indicate a duplication event that took place about 37 Mya.

Additionally, we have performed an alternative estimation of the duplication time based on a calculation of the evolutionary distance between two copies of the shared six transposable elements (see Materials and methods). The resulting distance value was 0.079. Assuming neutral evolution of the elements and taking the known substitution rate in primates, 2.2×10^{-9} changes per base pair per year (0.0022 changes per million years) [21], we calculated the duplication time to be about

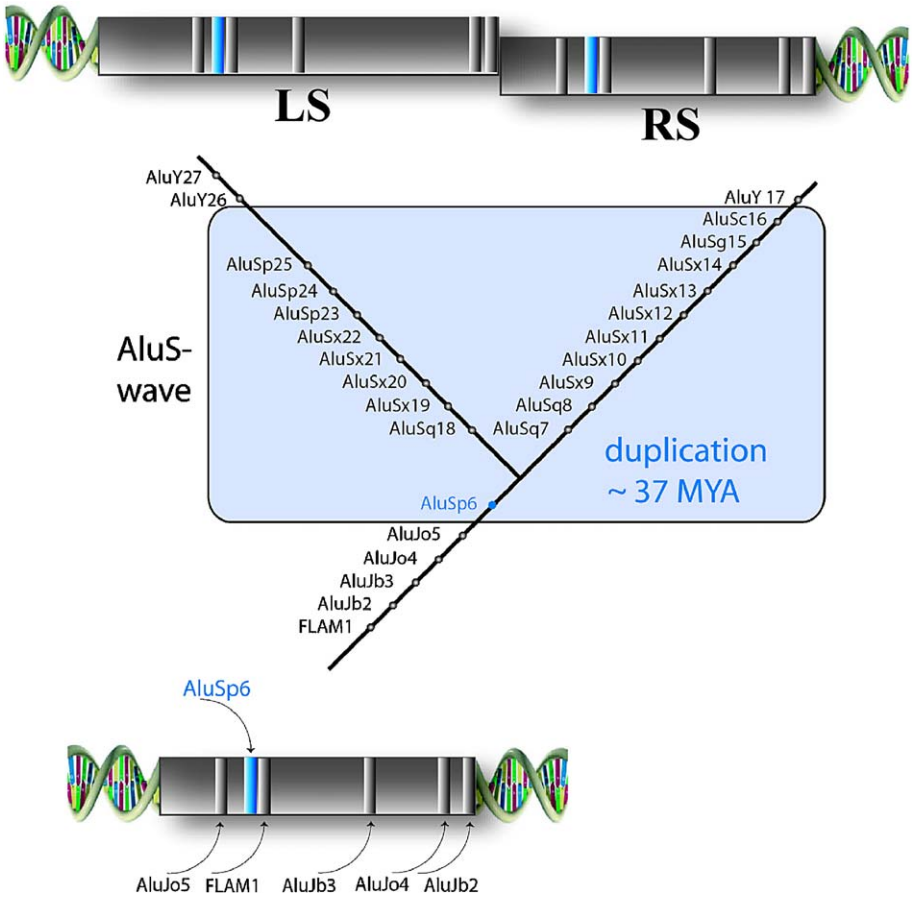

Fig. 1. Evolutionary time of duplication traced using *Alu* elements as a "molecular clock." The ancestor locus prior to the duplication is shown at the bottom. Six identical *Alu* elements were identified at corresponding positions in the left (LS) and right (RS) segments of this duplicated region. Among them the *AluSp6* element (blue) is about 37 My old, representing the oldest *Alu* element that inserted during the *AluS* wave and the youngest shared *Alu* element found in both duplicons. After duplication each duplicon accumulated additional *Alu* elements (*Alu* 7–17 for RS and *Alu* 18–27 for LS) independently.

36 Mya. Although this is a rough estimation, this value fits very well with the dating obtained by the insertion patterns of specific *Alu* elements.

The estimated time of the duplication coincides with the period when New World monkeys (NWM) diverged from the Old World monkey (OWM) lineage of the primates, about 35–40 Mya [22]. However, analysis of the *Alu* integrations is not sufficient to determine whether the segmental duplication occurred before or after the NWM split. To answer this question, we performed a Southern analysis of genomic DNA from six primate species, representing major primate clades: aye-aye and lemur—prosimians; squirrel monkey—NWM; guereza and gibbon—OWM; and human as hominid. The duplicated DNA fragments were detected using a probe specific for both segments. Two bands were observed in DNA samples of NWM as well as OWM and human (Fig. 2), suggesting that both segments are present in the corresponding genomes. By contrast, DNA samples of the prosimian species exhibit only single bands, suggesting a lack of the duplication in these animals. Therefore, we assume that the duplication took place in a primate lineage leading to anthropoid primates, after their separation from the prosimians but before the OWM and NWM split.

A further limit of the time estimation for the duplication event is based on an endogenous retrovirus sequence, MER51A, located in the LS. We found the orthologous MER51A sequences in the LS of the respective duplicated region in the *Macaca mulatta* genome (Table S2). This fact suggests that the MER51A insertion occurred after the duplication event but, at the latest, before the great apes split off from other OWM, about 25 Mya [22].

In summary, using independent approaches we show that the segmental duplication in 1q22 originated in a primate lineage leading to the anthropoids and occurred most probably about 36–37 Mya. Considering this coincidence, one may suppose that the SD1q22 locus provided new genetic material that may have contributed to the evolution of anthropoids. To identify potentially new genetic elements within the duplicated sequence, we compared this with its ancestral unduplicated form.

Reconstruction of the hypothetical unduplicated ancestor of SD1q22

To recover the original structure of the unduplicated proto-SD1q22 segment, we performed a detailed sequence analysis of

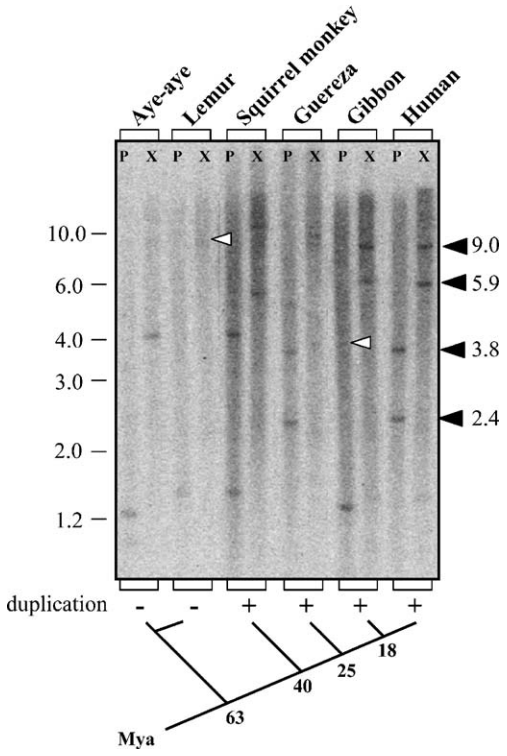

Fig. 2. Southern blot analysis of genomic DNA isolated from human and five other species of primates. Each sample of genomic DNA was digested separately with *Pvu*II (P) and *Xba*I (X). Sequence of the probe belongs to the human *GON4L* region on RS and exhibits 96.1% identity with the *YY1AP1* region on LS. White arrows indicate weak bands. Molecular markers (in kb) are indicated on the left. Band pairs representing LS and RS in human and their respective DNA fragment sizes (in kb) are indicated by black arrows. Evolutionary relationships among species are shown in the phylogenetic tree and the approximate times (Mya) of lineage splits are indicated (according to Goodman et al. [22]).

the left and right segments. We mapped the sequences using multiple computational and experimental approaches including gene and protein predictions, assembly and alignment of the cognate ESTs and mRNAs, identification and analysis of repetitive sequences, and cloning and sequencing of new cDNAs (see Materials and methods). Finally, we constructed the exact maps for both segments (Fig. 3 and supplementary UCSC Genome Browser custom track file). Comparative analysis of the LS and RS sequences revealed that prior to duplication the putative ancestral locus consisted of four protein-coding genes: *ASH1L, DAP3, misato*-like (*MSTO*), and a predicted gene of unknown function, which we call *gon4*-like, *GON4L*. *ASH1L* encodes a putative transcription factor that is the human ortholog of the *Drosophila* Absent, Small, or Homeotic discs 1 protein *ASH1* [23]. *DAP3* encodes a mitochondrial ribosomal protein, MRPS29, implicated in programmed cell death and also known as Death associated protein 3, DAP3 [24]. The *MSTO* gene is an ortholog of the essential cell division gene, *misato,* discovered in *Drosophila* [25,26]. The fourth gene, *GON4L,* has not yet been characterized and will be described below. This primary gene composition of the hypothetical proto-SD1q22 locus was also supported by analyses of the orthologous genomic loci in mouse, rat, and chicken, which lack this duplication (an exception being that the mouse locus contains an independent small duplication involving only the *Gon4l* gene; data not shown). The comparison revealed that the respective regions exhibit similar gene contents, gene order, and orientation (data not shown).

GON4L is a putative evolutionarily conserved transcription factor

To characterize the newly predicted *GON4L* gene, we first mapped and assembled the corresponding partial cDNA and EST sequences available in the database. Next, the compound sequences were verified by RT-PCR and extended by the RACE approach (see Materials and methods). As a result, we have determined the full coding region of this gene and its genomic

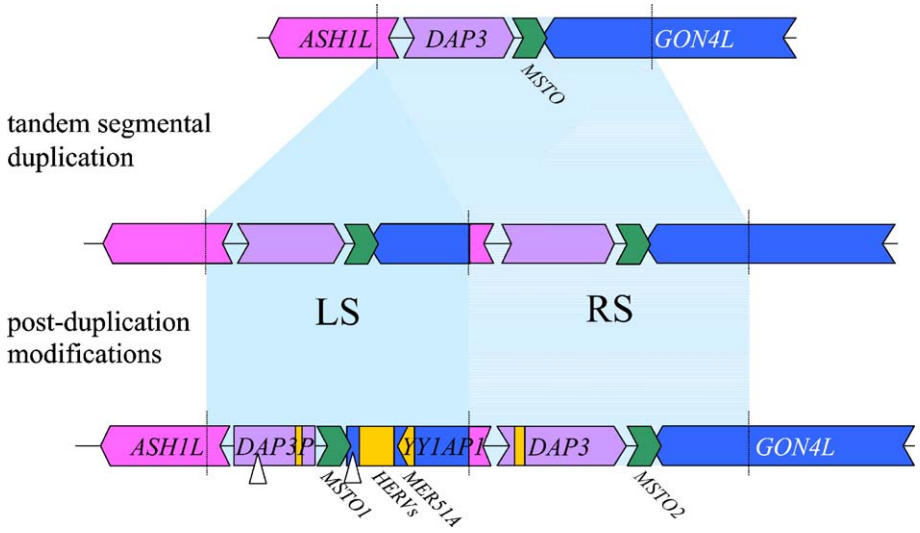

Fig. 3. Schematic representation of the proto-SD1q22 locus and its evolution. The hypothetical segment before duplication is shown at the top. General structure of the duplicated segments and the extant SD1q22 are presented in the middle and at the bottom, respectively. The locus orientation is shown with respect to the centromere (on the left). Insertions of retroviral elements are shown in yellow and large genomic deletions are indicated by arrowheads.

organization. In addition, we have analyzed its expression in various normal human tissues by Northern blot and quantitative RT-PCR. The *GON4L* gene is about 107 kb long (chromosome 1: 152532581–152640062) and consists of 32 exons. Expression of this gene is controlled by an alternative termination of transcription in intron 21, resulting in production of two alternative mRNAs. The full-length transcript, *GON4La* (AY335490), is 7676 bases long and encodes a 2241-amino-acid (aa) protein, while the truncated version, *GON4Lb* (AY335492), is 4755 bases long and encodes a 1529-aa protein. Both isoforms are ubiquitously expressed, although at varying levels in different tissues (Figs. 4A and 5). Furthermore, we have analyzed the protein coding region of *GON4La*. A search for sequence similarity in the protein databases (see Materials and methods) revealed that this gene encodes several evolutionarily conserved protein motifs that are shared by the orthologous genes in vertebrates, *Drosophila melanogaster, Caenorhabditis elegans, Dictyostelium discoideum,* and even plants (*Arabidopsis thaliana* and *Oryza sativa*) (Fig. S3). This suggests a very early origin of the *GON4L* gene, predating the radiation of plants and animals, and implies a function potentially essential for cell life. Among the orthologs of *GON4L,* we found only one gene that has been characterized at the functional level, *C. elegans gon-4,* which we used to name of the human ortholog. *gon-4* was identified as a cell lineage regulator of gonadogenesis in the worm [27]; however, no homologs of this gene had thus far been identified in other organisms. *gon-4* encodes a nuclear protein required directly for gonadogenesis in both sexes and, indirectly, for proper germline and vulval development. It was also concluded that *gon-4* may control expression of genes that drive the cell cycle. Another piece of evidence for the possible involvement of *GON4L* in cell cycle control comes from a study of the *Drosophila GON4L* ortholog, Cdp1. An interaction between Cdp1 and cyclin D (Cdi3) was demonstrated in a two-phase yeast two-hybrid screen [28]. In support of its expected function as a transcriptional factor, we found that all the orthologs of the *GON4L* gene, except those found in plants, encode a retrovirus-derived DNA-binding domain, known as SANT or Myb-like domain, PF00249 [29] (Fig. S3). Moreover, the three-dimensional structure of this domain of the mouse *GON4L* ortholog (known as hypothetical gene 2610100b20rik), bound to DNA, has been recently described in the protein data bank (PDB) (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1ug2). Taken together these data imply that *GON4L* likely encodes a transcription factor, whose possible function may be linked to cell cycle control.

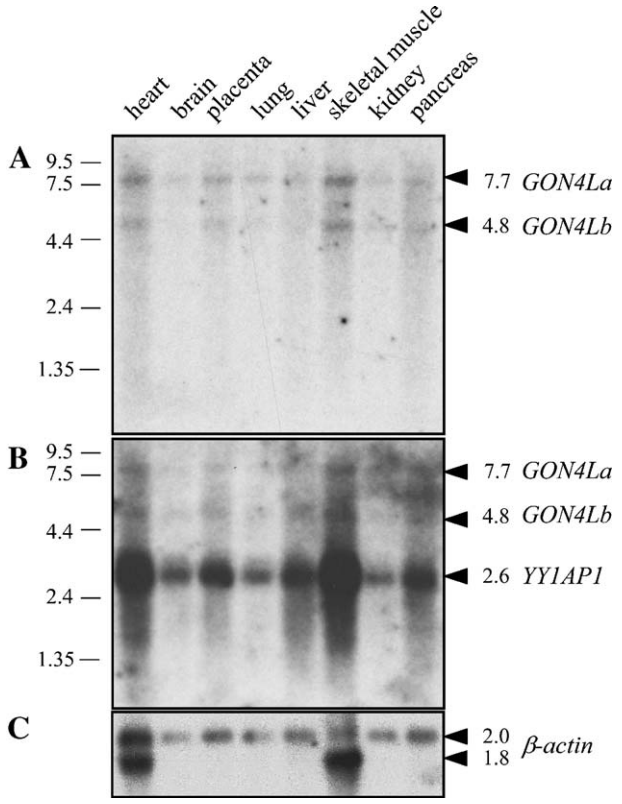

Fig. 4. Northern blot analysis of *GON4L* and *YY1AP1* in normal adult human tissues. The membrane was hybridized consecutively with three probes: (A) *GON4L* specific, (B) *YY1AP1/GON4L* specific, and (C) β-*Actin* specific. Molecular markers (in kb) are indicated on the left. The approximate sizes of transcripts are shown on the right (in kb).

Anthropoid-specific genetic material in SD1q22

When we compared the reference proto-SD1q22 segment with the extant LS and RS sequences of SD1q22, we identified additional genetic modifications that evolved after the initial duplication (Fig. 3). Initially, *DAP3* and *MSTO* each duplicated completely and gave rise to full-length paralogs, which subsequently evolved into a pseudogene and a truncated active gene, respectively. *ASH1L* and *GON4L* duplicated partially and gave birth to a new chimeric gene. Below, we provide detailed descriptions of these alterations.

Duplication of DAP3

DAP3, located in RS, has a fully preserved structure and currently represents the ancestral gene. Its copy in LS became a pseudogene, referred to here as *DAP3P,* which does not contain any meaningful open reading frame (data not shown). The respective genomic locus has undergone massive rearrangements including a large deletion and the insertion of several retroviruses and *Alu* retroposons (Fig. 3 and custom track file).

Duplication of MSTO

Misato was initially described in *Drosophila* as an essential cell division gene that encodes a protein with a partial similarity to tubulins [25]. Null mutations in the *misato* locus of *Drosophila* are associated with irregular chromosomal segregation during cell division. The orthologous gene in yeast, *Dml1* (*Drosophila misato*-like), was reported to play a role in mtDNA inheritance [30]. While the precise molecular function of *misato* remains unknown, its high degree of evolutionary conservation in eukaryotes (Fig. S4) suggests its possible importance for the cell.

We have determined the genomic organization of the paralogous *MSTO* genes, termed *MSTO1* and *MSTO2,*

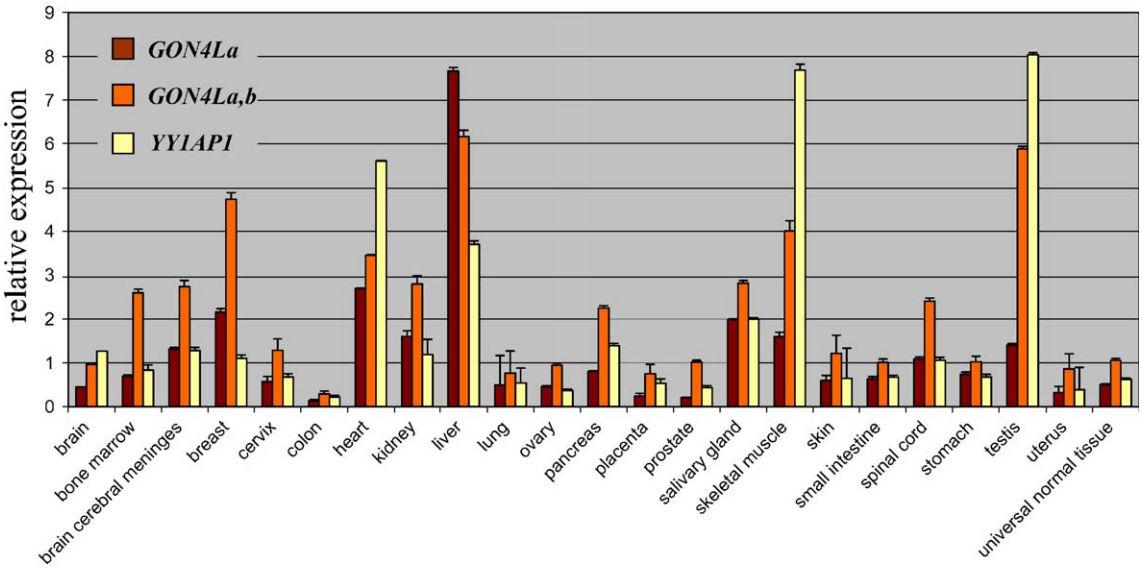

Fig. 5. Quantitative RT-PCR analysis of expression of *GON4La, GON4Lb,* and *YY1AP1* in adult human tissues. Signals obtained with a probe specific for *GON4La* are in brown, those from a dual-specific probe detecting both splice forms of *GON4L* (*GON4La, b*) are in orange, and signals from a *YY1AP1*-specific probe are in yellow. Bars represent standard error of the mean value.

according to their location in LS (chromosome 1: 152393080–152397830) and RS (chromosome 1: 152528660–152534001), respectively. The exon–intron structures of both genes are almost identical, with the exception that two *Alu* repeats were inserted into the intronic and 3′ untranslated regions of *MSTO2*. Both genes are transcriptionally active (Table S3). The exonic and intronic regions of *MSTO1* and *MSTO2* share the highest degree of nucleotide identity within the SD1q22 locus, 99.5 and 98.1%, respectively (Table S1). A comparative analysis of their protein coding regions revealed that *MSTO2* contains a dinucleotide deletion in exon 9 resulting in a reading frame shift, which leads to a C-terminal truncation immediately downstream of the third tubulin-related motif (Fig. S5 and Fig. 6). Preservation of this truncated isoform in human implies that it has a certain function that may compete with that of the main form.

We also analyzed the *MSTO1* and *MSTO2* genes in chimpanzee (panTro1), orangutan, and macaque. The respective genomic loci of the last two primates were assembled on the basis of available trace sequences (see Materials and methods). A schematic representation of the MSTO proteins is shown in Fig. 6. Each of these organisms retained only one intact copy of the *MSTO* gene, while the other copy was truncated. Considering that MSTO2 of chimpanzee, orangutan, and macaque exhibit differently truncated reading frames, which are shorter than human MSTO2, one may suppose that the truncation of the MSTO2 genes in these organisms occurred independently in a recent time period. The truncations in human and chimpanzee are not older than 7 My, while the one in orangutan is younger than 14 My. These data suggest that during evolution of anthropoids the *MSTO* gene, after its duplication, acquired dosage compensation independently in distinct primate lineages. On the other hand, only in human did the truncated form of *MSTO* retain the third tubulin-related motif and might, therefore, be functionally different, providing an example of a human-specific gene.

SD1q22 created a new anthropoid-specific gene

As a result of the segmental duplication, a new gene was formed in the fusion locus of the left and right segments juxtaposing the 5′ region of *ASH1L* and 3′ region of *GON4L*

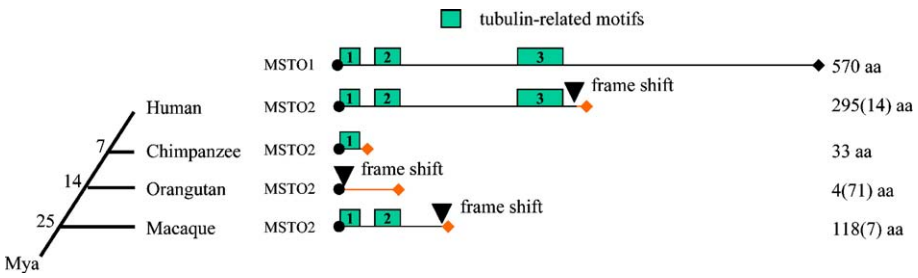

Fig. 6. Schematic representation of the MSTO1 and MSTO2 proteins in primates. MSTO1 is given as a reference sequence. The approximate times of the lineage splits (Mya) are indicated on the left (according to Goodman et al. [22]). Closed circles represent the start methionines, terminal residues are indicated by rhomboids, and arrowheads indicate frameshift mutations. In the orangutan this occurs at residue 4 of the protein and changes the 67 residues of the putative protein sequence downstream (indicated in orange). Tubulin-related motifs in the protein sequences are indicated by green boxes numbered 1–3.

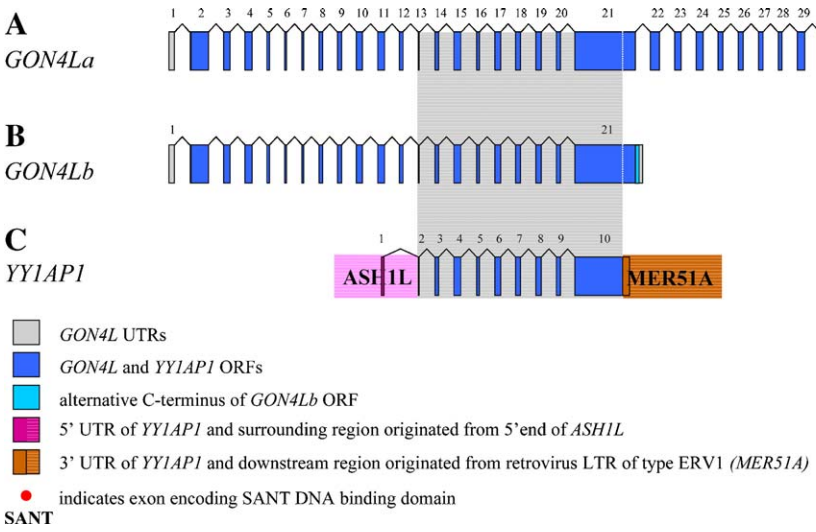

Fig. 7. Schematic representation of the genomic organization of *GON4L* and *YY1AP1*. Exons are indicated by scaled rectangles. ORFs are shown in blue. (A) Isoform *GON4La* and (B) isoform *GON4Lb*. Exon 21 of *GON4Lb* contains the alternative C-terminus indicated in turquoise. Exon 31 encodes the SANT DNA-binding domain (indicated by red circle). (C) Structure of the *YY1AP1* gene.

(Fig. 3). Our comparative sequence analysis revealed that the new fusion gene consists of the partial promoter and untranslated first exon of *ASH1L* and coding exons 13–21 of *GON4L* (Fig. 7). Termination of transcription and translation of this gene occur on the long terminal repeat (LTR) of an endogenous retrovirus, also known as the MER51A repeat, that was inserted into exon 21 of *GON4L* after the duplication, once more documenting contribution of retroposons to gene architecture [31]. Expression products of this gene were previously described in two independent studies as hepatocellular-associated protein 2, HCCA2 [32], or as a coactivator of the zinc finger transcription factor YY1, YY1-associated protein (YY1AP1) [33]. Protein comparison of GON4L and YY1AP1 shows that 748 of 750 aa residues of YY1AP1 correspond to aa residues 590–1337 of the GON4L protein. The last 2 C-terminal residues of YY1AP1 originate from the LTR of MER51A. Both proteins exhibit a high degree of sequence similarity, suggesting that the YY1AP1 protein preserved function of the respective protein region of GON4L and, as such, may represent a truncated version that is similar to GON4Lb minus the SANT DNA-binding domain (Fig. 7). The high sequence similarity also implies that if both proteins are coexpressed in the same cell, then YY1AP1 may interfere with the function of GON4L. We analyzed the expression of *YY1AP1* relative to both isoforms of *GON4L* in different human tissues by Northern blot (Fig. 4) and quantitative RT-PCR (Fig. 5). Similar to *GON4L, YY1AP1* is also expressed ubiquitously except that its highest levels occur in the testis and skeletal muscle, suggesting that this gene may be involved in protein–protein interactions, competing with both isoforms of *GON4L*. The fact that YY1AP1 is an anthropoid-specific transcriptional cofactor highly expressed in testis also implies that the new activity of this gene could be associated with anthropoid speciation.

In conclusion, our data indicate that SD1q22 created anthropoid-specific genetic material whose new functions could have contributed to speciation of the anthropoid primates. The detailed information we have reported regarding the structure and evolution of this region will provide a solid foundation for future studies of its anthropoid-specific functions and possible linkage to human genetic disorders.

Materials and methods

Sequence analyses

The UCSC Genome Browser (http://genome.ucsc.edu/) and NCBI Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) served as the main sources of genomic, transcript, and protein sequences in our study. We used the following genome assemblies: human—hg17; chimpanzee—panTro1; macaque—rheMac1; mouse—mm6; rat—rn3; chicken—galGal2; zebrafish—DanRer2; fruit fly—dm2. The UCSC Genome Browser was used to map cDNA sequences and to extract large genomic sequences.

We used the Advanced PipMaker [34] for comparative analyses of large regions of genomic DNA and the RepeatMasker (http://www.repeatmasker.org) to map locations of repetitive elements. Alignments of genomic segments, transcripts, and proteins were carried out using CLUSTALW [35] and DIALIGN2 [36], and the resulting alignments were adjusted manually in GeneDoc (http://psc.edu/biomed/genedoc). In the calculation of sequence identity between paralogous regions we excluded repeats and sites that contained insertions or deletions. The GENSCAN [37] and TWINSCAN [38] programs were used to predict the potential gene structure of *GON4L,* which was then improved manually. Trace sequences related to MER51A insertion in primates were obtained from the NCBI Trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) and assembled into contigs using SeqMan (DNASTAR). We defined subtypes of the *Alu* repeats and their ages according to Kapitonov and Jurka [20].

Protein similarity was analyzed using PSI-BLAST (NCBI) [39] and MOTIF SEARCH (http://motif.genome.jp) with protein databases available at NCBI. The SMART system was used for protein domain identification [40]. All human sequence data used in this study are included in the custom track file (http://www.dkfz.de/mga/home/SD1q22/customtracks.txt).

Sequence-based dating

We concatenated the shared six orthologous retroposed elements and assembled an alignment including the sequences of both duplicated segments.

Gap positions and variable CpG sites were removed. The remaining 1346 nt comprise 106 variable positions. The sequence-based dating was performed by assuming an average mammalian genome mutation rate of 2.2×10^{-9} per base pair per year [21].

RT-PCR and RACE analysis of the GON4L transcripts

Total RNA for RT-PCR was isolated from normal human tissues (colon, brain, heart, ovary, pancreas, and skin) kindly provided by Prof. Dr. med László Füzesi (Center for Pathology, Göttingen, Germany). Isolation of RNA, cDNA synthesis, and RT-PCR were performed as previously described [41]. 5′- and 3′-RACE experiments were carried out using total RNA from normal human skin (Invitrogen) and the BD SMART RACE cDNA Amplification Kit (BD Biosciences Clontech). The RACE PCR products were cloned via TA cloning (pGEM-T Easy Vector System; Promega) and sequenced with the BigDye chemistry and ABI Prism 3100 sequencers (Applied Biosystems). Sequences were assembled into contigs as described above.

Southern blot

Human genomic DNA was kindly provided by Ben Schütul (DKFZ, Germany). DNA from five species of primates (aye-aye (*Daubentonia madagascariensis*), lemur (*Lemur catta*), squirrel (*Saimiri sciureus*), guereza (*Colobus quereza*), and gibbon (*Hylobates lar*)) was isolated by the standard phenol/chloroform method [42] from tissues kindly provided by Christian Roos (German Primate Center, Göttingen). Ten micrograms of each DNA sample was digested with *Pvu*II or *Xba*I (Fermentas), fractionated in a 0.7% agarose gel, and immobilized on a nylon membrane (Hybond N+; Amersham) by the standard alkaline transfer method [42]. A probe with dual specificity for the *GON4L* and *YY1AP1* genes was generated by PCR from cDNA clone DKFZP686J14104 (CR749789) using the primer pair 5′-CCTGAAACTGAAGCCAGTTGCCAC and 5′-GGAAGTAGGGTTAACCAAGAGGGTAGTG and labeled with [α-^{32}P]dCTP by random priming using the HexaLabel Plus DNA Labeling Kit (Fermentas) following the manufacturer's protocol. The membrane was hybridized overnight at 65°C in Church buffer (0.5 M NaP_i, pH 7.2, 7% SDS, 1 mM EDTA, pH 8) and then washed under high-stringency conditions (1× SSC, 0.1% SDS at 65°C).

Northern blot

A human multiple tissue Northern blot (MTN) was purchased from Clontech (Human MTN Blot; Clontech). *GON4L*- and *YY1AP1*/*GON4L*-specific probes were amplified by PCR from clones DKFZp686H0793 (BX537764) and DKFZP686J14104 (CR749789), respectively. Primers used for the amplification of a *GON4L*-specific product were 5′-ATGCAGGAAATCAGCTTGGTATGGAG and 5′-AGTCCGGAAATCCTCTGTGTCTGG. A probe matching both *YY1AP1* and *GON4L* transcripts was amplified with primers 5′-AATCTCAATCCGGAGGCCAGTAG and 5′-CCCAGTCTCCAGACTCTTATTCTCCTAGCTCAAAG. DNA probes were [α-^{32}P]dCTP labeled with a random-priming kit (Boehringer Mannheim). Hybridization was carried out overnight at 65°C in Church buffer followed by washing under high-stringency conditions (1× SSC, 0.1% SDS at 65°C). Membranes were stripped in boiling 0.5% SDS for 30 min under vigorous shaking.

Quantitative real-time RT-PCR

Total RNA (50 μg), extracted from various human tissues (BioCat and BD Bioscience), was reverse transcribed using the RevertAid H Minus First Strand cDNA Synthesis Kit (Fermentas). Fluorescent TaqMan probes and gene-specific primers were obtained from Applied Biosystems (primer sequences are available upon request). The assays were performed using an ABI Prism 7900 HT sequence detector (Applied Biosystems). Thermal cycling conditions were 94°C for 15 s, 56°C for 30 s, and 76°C for 30 s, 45 cycles. Each sample was analyzed in triplicate and data analysis was carried out with the SDS 2.1 software (Applied Biosystems). The relative quantification of mRNA expression levels of genes was estimated using the comparative C_T method ($2^{-\Delta\Delta CT}$) [43]. β-Actin was used as an internal control and RNA from a cell line pool (Universal Human Reference RNA; Stratagene) was taken as a reference for the expression analysis.

Acknowledgments

The authors thank Jürgen Brosius for valuable comments on the manuscript; Markus Seiler and Carsten Raabe for helpful discussions during data analysis; Anny Duda, Anja Kolb, and Ursula Jordan for technical assistance; Marsha Bundman for editorial help. This work was supported by the Bundesministerium für Bildung und Forschung (Grants 01GR0101 and 01GR0420, National Genome Research Network). J.S. and T.S.R. were supported in part by a grant from the Nationales Genomforschungsnetz (0313358A to Jürgen Brosius).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.02.002.

References

[1] S. Ohno, Evolution by Gene Duplication, Springer-Verlag, New York, 1970.
[2] M. Kimura, T. Ota, On some principles governing molecular evolution, Proc. Natl. Acad. Sci. USA 71 (1974) 2848–2852.
[3] E.E. Eichler, Recent duplication, domain accretion and the dynamic mutation of the human genome, Trends Genet. 17 (2001) 661–669.
[4] R.V. Samonte, E.E. Eichler, Segmental duplications and the evolution of the primate genome, Nat. Rev. Genet. 3 (2002) 65–72.
[5] J.A. Bailey, et al., Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22, Am. J. Hum. Genet. 70 (2002) 83–100.
[6] H. Jin, et al., Structural evolution of the BRCA1 genomic region in primates, Genomics 84 (2004) 1071–1082.
[7] Y. Ji, E.E. Eichler, S. Schwartz, R.D. Nicholls, Structure of chromosomal duplicons and their role in mediating human genomic disorders, Genome Res. 10 (2000) 597–610.
[8] R. Mazzarella, D. Schlessinger, Pathological consequences of sequence duplications in the human genome, Genome Res. 8 (1998) 1007–1021.
[9] C.J. Shaw, J.R. Lupski, Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease, Hum. Mol. Genet. 1 (Spec. No. 1) (2004) R57–R64.
[10] P. Stankiewicz, et al., Genomic disorders: genome architecture results in susceptibility to DNA rearrangements causing common human traits, Cold Spring Harbor Symp. Quant. Biol. 68 (2003) 445–454.
[11] S. Wiemann, et al., Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs, Genome Res. 11 (2001) 422–435.
[12] S. Wiemann, S. Bechtel, D. Bannasch, R. Pepperkok, A. Poustka, The German cDNA network: cDNAs, functional genomics and proteomics, J. Struct. Funct. Genom. 4 (2003) 87–96.
[13] K.W. Cheng, et al., The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers, Nat. Med. 10 (2004) 1251–1256.
[14] S. Knuutila, Cytogenetics and molecular pathology in cancer diagnostics, Ann. Med. 36 (2004) 162–171.
[15] N. Mandahl, F. Mertens, I. Panagopoulos, S. Knuutila, Genetic characterization of bone and soft tissue tumors, Acta Orthop. Scand. Suppl. 75 (2004) 21–28.
[16] N. Wong, et al., Positional mapping for amplified DNA sequences on 1q21–q22 in hepatocellular carcinoma indicates candidate genes over-expression, J. Hepatol. 38 (2003) 298–306.
[17] J. Jurka, O. Kohany, A. Pavlicek, V.V. Kapitonov, M.V. Jurka, Duplication,

coclustering, and selection of human Alu retrotransposons, Proc. Natl. Acad. Sci. USA 101 (2004) 1268–1272.
[18] J.A. Bailey, G. Liu, E.E. Eichler, An Alu transposition model for the origin and expansion of human segmental duplications, Am. J. Hum. Genet. 73 (2003) 823–834.
[19] D.H. Kass, M.A. Batzer, P.L. Deininger, Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution, Mol. Cell. Biol. 15 (1995) 19–25.
[20] V. Kapitonov, J. Jurka, The age of Alu subfamilies, J. Mol. Evol. 42 (1996) 59–65.
[21] S. Kumar, S. Subramanian, Mutation rates in mammalian genomes, Proc. Natl. Acad. Sci. USA 99 (2002) 803–808.
[22] M. Goodman, et al., Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence, Mol. Phylogenet. Evol. 9 (1998) 585–598.
[23] T. Nakamura, et al., huASH1 protein, a putative transcription factor encoded by a human homologue of the Drosophila ash1 gene, localizes to both nuclei and cell–cell tight junctions, Proc. Natl. Acad. Sci. USA 97 (2000) 7284–7289.
[24] J.L. Kissil, et al., Isolation of DAP3, a novel mediator of interferon-gamma-induced cell death, J. Biol. Chem. 270 (1995) 27932–27936.
[25] G.L. Miklos, M. Yamamoto, R.G. Burns, R. Maleszka, An essential cell division gene of Drosophila, absent from Saccharomyces, encodes an unusual protein with tubulin-like and myosin-like peptide motifs, Proc. Natl. Acad. Sci. USA 94 (1997) 5189–5194.
[26] R.L. Strausberg, et al., Generation initial analysis of more than 15,000 full-length human and mouse cDNA sequences, Proc. Natl. Acad. Sci. USA 99 (2002) 16899–16903.
[27] L. Friedman, S. Santa Anna-Arriola, J. Hodgkin, J. Kimble, gon-4, a cell lineage regulator required for gonadogenesis in *Caenorhabditis elegans*, Dev. Biol. 228 (2000) 350–362.
[28] J. Zhong, H. Zhang, C.A. Stanyon, G. Tromp, R.L. Finley Jr., A strategy for constructing large protein interaction maps using the yeast two-hybrid system: regulated expression arrays and two-phase mating, Genome Res. 13 (2003) 2691–2699.
[29] R. Aasland, A.F. Stewart, T. Gibson, The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB, Trends Biochem. Sci. 21 (1996) 87–88.
[30] A. Gurvitz, A. Hartig, H. Ruis, B. Hamilton, H.G. de Couet, Preliminary characterisation of DML1, an essential Saccharomyces cerevisiae gene related to misato of Drosophila melanogaster, FEMS Yeast Res. 2 (2002) 123–135.
[31] J. Brosius, S.J. Gould, On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA", Proc. Natl. Acad. Sci. USA 89 (1992) 10706–10710.
[32] Z.X. Wang, H.Y. Wang, M.C. Wu, Identification and characterization of a novel human hepatocellular carcinoma-associated gene, Br. J. Cancer 85 (2001) 1162–1167.
[33] C.Y. Wang, et al., YY1AP, a novel co-activator of YY1, J. Biol. Chem. 279 (2004) 17750–17755.
[34] S. Schwartz, et al., PipMaker—A Web server for aligning two genomic DNA sequences, Genome Res. 10 (2000) 577–586.
[35] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[36] B. Morgenstern, DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment, Bioinformatics 15 (1999) 211–218.
[37] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.
[38] I. Korf, P. Flicek, D. Duan, M.R. Brent, Integrating genomic homology into gene structure prediction, Bioinformatics 17 (Suppl. 1) (2001) S140–S148.
[39] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[40] I. Letunic, et al., SMART 4.0: towards genomic data integration, Nucleic Acids Res. 32 (2004) D142–D144 (Database issue).
[41] S. Gupta, D. Zink, B. Korn, M. Vingron, S.A. Haas, Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing, BMC Genom. 5 (2004) 72.
[42] J. Sambrook, E. Fritsch, T. Maniatis, Molecular Cloning: a Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989.
[43] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2(-\Delta\Delta T)$ method, Methods 25 (2001) 402–408.